

---

# Reprocessing Study

## Nicholas Singer

[nsinger@eos.hitc.com](mailto:nsinger@eos.hitc.com)

---

31 October 1995



# Overview

**Meeting reprocessing requirements is a significant driver of hardware purchases and designs:**

- CPU power required
- Data set organization within archive (on physical media)
- Archive needs (volume-driven vs. transaction-driven)
- Communications and I/O bandwidth requirements
- Error detection and control approach

**We'll cover**

- **How Do We Characterize Reprocessing?**
  - What kind and quantity of processing it entails
  - What "Reprocess at 2x rate" might mean
- **Planned Modeling Studies**

**ECS Context**

- **Data Server**
- **Networks**
- **Processors**



# Requirements for Reprocessing

**EOSD1040: “ECS shall provide sufficient capacity to permit reprocessing of all EOS science data at twice the incoming data rate at a minimum, concurrently with processing of new data.”**

**August 1995 *Technical Baseline for the ECS Project*, Attachment L, provides a phasing of required processing capacities, relative to launch. For epoch k, 3Q99, this is an additional 1x for AM-1 instruments, an additional 2x for TRMM instruments.**

***Problem*—This doesn’t necessarily match the future reprocessing needs of the instrument teams.**

***Problem*—Uncertainties about reprocessing paradigms and reprocessing frequencies lead to large uncertainties in the hardware requirements.**

# Approaches to Quantifying Capacities for Reprocessing



**Basic:** Take capacities required to meet near-peak\* 1x requirements for current processing and multiply them by 2.

**Better:** Model reprocessing loads (@1x or 2x current) mixed with current processing loads; give reprocessing a lower priority than current processing; calculate required capacities.

**Best:** Use reprocessing plans from instrument teams to model reprocessing load; add 1x current load; model dynamically.

\* “Near-peak requirements” means capacity needed to process peak loads within time period allowed (e.g., 24 hours for Levels 1-3.)

# Comparison of 1x “Push” Processing Requirements for LaRC (Nominal MFLOPS)

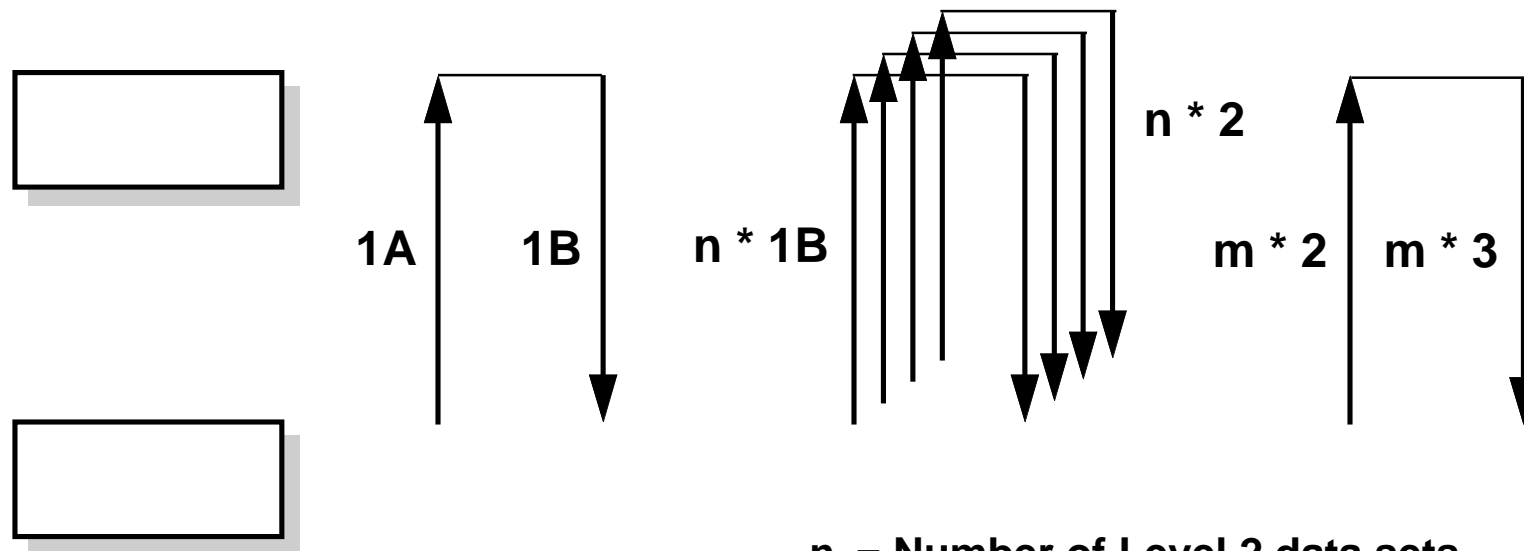
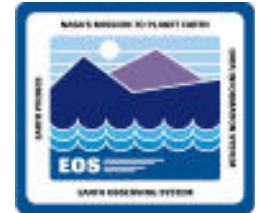


August 1995 Baseline; Epoch k (3Q99)	Average 1x (Static)	Busy-Day 1x (Static)	Near-Peak 1x (Dynamic)	Double [Triple for TRMM] Near-Peak 1x	Near-Peak 1x + Average 1x [+ another 1x for TRMM]
<b>CERES TRMM</b>	3,071	3,298	4,320	12,960	11,520
<b>CERES AM</b>	6,825	7,423	8,640	17,280	15,840
<b>MISR</b>	13,644	13,645	14,400	28,800	28,800
<b>MOPITT</b>	26	27	<1 processor	<1 processor	<1 processor

*Sizing is rounded up to the next pair of 720 MFLOPS processors.*



# PGS – Archive Traffic for Reprocessing Paradigm



$n$  = Number of Level 2 data sets

$m$  = Number of Level 3 data sets

*Individual Product*

# Effects of Reprocessing Paradigms on LAN Flows



***Relative sustained Archive-to-PGS LAN bandwidth required to reprocess 1 day in 1 day***

	Head of Chain	Level Contemporaneous	Individual Product
LaRC TRMM	1.0	1.9	3.2
LaRC AM-1	19.5 *1.0	56.3 *2.9	155.1 *8.0

***\* Renormalized numbers***

# Theoretical Models of Reprocessing Frequency



**Constant Interval – The period between reprocessings is constant**

- Data set will be reprocessed  $N$  times each year.

**Linear Interval – Interval between reprocessings increases at fixed rate**

- Interval determined by time since launch

**Constant Decay – Interval between reprocessings increases at variable rate**

- Interval determined by algorithm delivery number

**Logistics Model (B. Barkstrom)**

- Reprocessing load in first campaign derived from logistic equation

**Other models are possible, but were not examined**

- Since reprocessing drivers are complex, behavior can be complex

# Summary of Reprocessing Frequency Models



	<b>Hardware Needs</b>	<b>Reprocessing Interval</b>	<b>Comments</b>
<b>Constant Interval</b>	Increasing	Constant	
<b>Linear Interval</b>	Constant	Increasing	<b>Fixed Resources</b>
<b>Constant Decay</b>	<b>Compromise (Increasing)</b>	<b>Compromise (Increasing)</b>	<b>Increasing Knowledge</b>



# Current Status

## Status

- IDR modeling assumes head-of-chain reprocessing paradigm
- 2x can be estimated as 1x peak + 1x average
- 3x can be estimated as 1x peak + 2\*(1x average)

## Critical Decisions

- With AHWGP and instrument teams, decide how to quantify the 2x requirement
  - Reprocessing paradigm(s)
  - Reprocessing frequency

## Next Steps

- Validate reprocessing requirements with AHWGP, instrument teams
- Model accordingly for CDR



# Planned Modeling Studies

**In process now, making model runs to contrast the following assumptions (for Release B, Epoch k):**

- **Double the near-peak processing requirements for 1x**
- **Take the near-peak dynamic processing requirements and add 1x average (static) processing requirements**
- **Using the head-of-chain paradigm in the dynamic model, double the frequency and halve the coverage of each product/process and calculate near-peak processing requirements**
- **Do the same, but let one instantiation have high priority (“Current processing”) and one with low priority (“Reprocessing”)**
- **Using the head-of-chain paradigm in the dynamic model, create two copies of L1-L4 processing. Let one copy have high priority (“Current processing”) and one have low priority (“Reprocessing”)**

**In the future, contrast with**

- **Explicit, time-phased reprocessing scenarios from the instrument teams (paradigms other than head-of-chain)**

# Summary



**Reprocessing assumptions are a significant driver of hardware capacity requirements**

**Different forms of reprocessing have different effects on networks, data server, etc.**

**The basic approach of multiplying 1x near-peak processing capacities by a factor does not properly account for the effects of different reprocessing paradigms**